### Thanks to:

- NHLBI Grant: HL089102
- Rajiv Dulepet
- Michael Angerman
  (Constantin Georgescu)
- Ellen Rothenberg & lab
- Barbara Wold & lab
  (Caltech)
- Guy Naor, Cosmin Andriescu
  (SparkTech Soft)
- Martin Morgan (FHCRC)

### High throughput molecular and cell biology

#### Technology characteristics

<table>
<thead>
<tr>
<th>Technology characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small sample sizes</td>
</tr>
<tr>
<td>Commoditized equipment</td>
</tr>
<tr>
<td>Low-cost per experiment</td>
</tr>
<tr>
<td>Global (genome-wide) data</td>
</tr>
<tr>
<td>Rapidly evolving</td>
</tr>
<tr>
<td>Currently dominated by</td>
</tr>
<tr>
<td>- Re-sequencing</td>
</tr>
<tr>
<td>- ChiP-seq</td>
</tr>
<tr>
<td>- RNA-seq</td>
</tr>
<tr>
<td>But also</td>
</tr>
<tr>
<td>- Proteomics, Metabolomics</td>
</tr>
<tr>
<td>- Screens (RNAI, drugs, mutants)</td>
</tr>
<tr>
<td>- Biomarker assays</td>
</tr>
<tr>
<td>- Single-cell assays</td>
</tr>
<tr>
<td>(microfluidics, FACS, ...)</td>
</tr>
</tbody>
</table>

#### Computational implications

<table>
<thead>
<tr>
<th>Computational implications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large, noisy datasets</td>
</tr>
<tr>
<td>- Complex analysis needed</td>
</tr>
<tr>
<td>- Results are statistical &amp; parameter/method dependent</td>
</tr>
<tr>
<td>Accessible to individual labs</td>
</tr>
<tr>
<td>- Global data, focused hypotheses</td>
</tr>
<tr>
<td>- Bursts of heavy computing</td>
</tr>
<tr>
<td>- Opportunity: iterative data exploration</td>
</tr>
<tr>
<td>Large-scale cataloguing projects</td>
</tr>
<tr>
<td>- (TGCA, ENCODE, Epigenome, 1000 genomes, Bacteriome...)</td>
</tr>
<tr>
<td>- Need to share data + analysis</td>
</tr>
<tr>
<td>- Opportunity:</td>
</tr>
</tbody>
</table>
Genes make RNAs, which make proteins, some of which regulate gene expression
ChIP-seq Overview

1. Cross-link proteins to DNA
2. Fragment chromatin to 100-150bp
3. Immunoprecipitate antibody-bound DNA fragments
4. Reverse cross-links and sequence fragment ends
5. Map sequence reads to genome
6. Identify genomic regions with enriched number of mapped reads

A typical pipeline:
(Hes1 ChIP-seq in HSCs Bernstein lab, FHCRC)
> 500 lines of R
Uses 12 R “packages”
Includes > 10 major user-defined parameters
e.g. # read mismatches, Baye’s PP, ncRNAs?
> 200 interim files

BaysPeak: Bayesian analysis of ChIP-seq data
(T Pinterest Score: 14.8. cigar plot and summary: Target)

Barbara Wold’s lab, Science, 2007, 16(5830):1497-502
Two types of complexity in high throughput data processing – (1) difficult math

\[ l(\Theta|y) \] is the complete-data log-likelihood, given by

\[
l(\Theta|y) = \sum_{d \in \{(x_r)\}} \sum_{t=1}^T z_{nt} \left\{ \log \left[ w_n \mathcal{N} \left( d_j | \mu_{nt}, \frac{1}{\sigma_{nt}} \right) \mathcal{G}(u_{nt}|2, 2) \right] \right\} \]

\[
= \sum_{d \in \{(x_r)\}} \sum_{t=1}^T z_{nt} \left\{ \log w_n - \log \sigma_n - \log \sqrt{2\pi} - \frac{1}{2} \left( \frac{d_j - \mu_{nt}}{\sigma_{nt}} \right)^2 + \log u_{nt} - 2\log + \log 1 \right\}
\]

and \( f_{prior} \), the log prior ‘penalty’ on \((\delta, \sigma^2, \sigma)\), is given as

\[
f_{prior} = -\frac{1}{2} \sum_x \left\{ (\sigma^2 + \sigma^2)^{2} | x(x - \xi)^2 + \beta \right\} + \frac{2\alpha - 1}{2} \sum_x \left\{ \log(\sigma^2 + \sigma^2) \right\}. \quad (6)
\]

E-step: Given the current estimate \( \Theta^{(t)} \) for \( \Theta \), the conditional expectation of the penalized log complete data likelihood is given as

\[
Q(\Theta; \Theta^{(t)}) = \mathbb{E}[l(\Theta|y)|y] + f_{prior}
\]

\[
= \sum_{d \in \{(x_r)\}} \sum_{t=1}^T z_{nt} \left\{ \log w_n - \log \sigma_n - \log \frac{\hat{\sigma}_n}{\sigma_n} \right\} + \lambda \quad (7)
\]

where \( \lambda \) is a constant with respect to the parameter vector \( \Theta \). Given this, the E-step [Mclachlan 2000] consists of computing the following quantities

\[
\hat{z}_{nt} \triangleq \mathbb{E}(Z_{nt}|y, \Theta^{(t)}) = \frac{\pi w_n (d_j | \mu_{nt}, \sigma_{nt})}{\sum_z w_n (d_j | \mu_{nt}, \sigma_{nt})}
\]

\[
\hat{e}_{nt} \triangleq \mathbb{E}(E_{nt}|y, \Theta^{(t)} = 1, \Theta^{(t)}) = \frac{\tilde{\sigma}_n}{\hat{\sigma}_n} \quad (8)
\]

Two types of complexity in high throughput data processing – (2) difficult choices

e.g. for ChIP-seq:

- Average fragment length
- Scan window size
- Read count significance threshold
- Background reads distribution
- Sensitivity vs. specificity

![Graph showing sensitivity vs. FAR](image-url)
N1 Hes1

DNA fragments (excluding 120bp adapters) are asymmetrically distributed ~ 50bp–280bp
(data: Suzanne Furuyama, Bernstein lab, FHCRC)

Torres, Metta, Ottenwälder & Schlötterer, Genome Research, 2007, 18:000

Kharchenko, Tolstorukov and Park

Estimating the average DNA fragment length from the strand-specific tag shift

Sarkar, Gentleman, Lawrence, Zhang, Yao

Kharchenko, Tolstorukov and Park


Assuming average fragment length=75bp

Assuming average fragment length=136bp
Ly9 expression is repressed in the bone marrow.

Data Suzanne Furuyama & Irwin Bernstein, FHCRC
<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Control?</th>
<th>Parameters</th>
<th>strands?</th>
<th>Map?</th>
<th>Significance score</th>
<th>FDR measure?</th>
<th>TF/histones?</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChIP</td>
<td>control or IP only</td>
<td>read length</td>
<td>merge strands</td>
<td>no shift</td>
<td>N</td>
<td>simple height</td>
<td>empirical</td>
</tr>
<tr>
<td>XSET</td>
<td>IP only</td>
<td>fragment length orientation</td>
<td>merge strands</td>
<td>no shift</td>
<td>Y</td>
<td>simple height</td>
<td>empirical</td>
</tr>
<tr>
<td>Mikkelsen et al</td>
<td>IP only</td>
<td>no orientation</td>
<td>no merge</td>
<td>no shift</td>
<td>Y</td>
<td>p-values from permutations</td>
<td>no official FDR</td>
</tr>
<tr>
<td>MACS</td>
<td>control or IP only</td>
<td>fragment length orientation</td>
<td>no duplicated reads</td>
<td>merge strands</td>
<td>N</td>
<td>Poisson values</td>
<td>FDR by peaks in control/IP</td>
</tr>
<tr>
<td>QUEST</td>
<td>control</td>
<td>orientation</td>
<td>shift reads</td>
<td>merge strands</td>
<td>N</td>
<td>kernel density estimation</td>
<td>FDR by permutations of the control</td>
</tr>
<tr>
<td>FastPeaks</td>
<td>IP only</td>
<td>fragment length orientation</td>
<td>no merge</td>
<td>no shift</td>
<td>N</td>
<td>Poisson values</td>
<td>FDR by permutations of the IP</td>
</tr>
<tr>
<td>SISRA</td>
<td>control or IP only</td>
<td>fragment length orientation</td>
<td>no merge</td>
<td>no shift</td>
<td>N</td>
<td>compare reads on different strands</td>
<td>FDR by peaks in background/IP</td>
</tr>
<tr>
<td>Kharatchkin et al</td>
<td>control</td>
<td>orientation</td>
<td>no merge</td>
<td>no shift</td>
<td>N</td>
<td>Poisson distribution</td>
<td>FDR by permutations of the control</td>
</tr>
<tr>
<td>PeakSeq</td>
<td>control</td>
<td>fragment length orientation</td>
<td>merge strands</td>
<td>Y</td>
<td>sample normalization Binomial distribution</td>
<td>FDR estimate, q-values (BH correction)</td>
<td>both</td>
</tr>
<tr>
<td>BayesPeak</td>
<td>control or IP only</td>
<td>fragment length orientation</td>
<td>no merge</td>
<td>no shift</td>
<td>N</td>
<td>negative binomial distribution, Bayesian posterior probabilities</td>
<td>FDR by enrichment probabilities</td>
</tr>
</tbody>
</table>

Spyrou et al, BMC Bioinformatics 2009, 10:299

GABP

FOXA1

Zhang et al, Biometrics 2010, Jun 1st [Epub ahead of print]
NIH Roadmap Epigenomics consortium

292 public datasets so far

Order of magnitude estimates of biological data sizes

- Read data for 1 human genome ~ 250 GB
- Annotated variants for 1 human genome ~ 25 GB
- 1000 genomes ~ 25 TB
- ChIP-seq read data for 1 TF in 1 condition ~ 1.5 GB
- ChIP-seq read data for 2,500 TFs in 200 cell types ~ 750 TB
- RNA-seq read data for 1 gene (ELAND format) ~ 500 MB
- RNA-seq read data for 22,000 genes in 200 cell types ~ 2.2 PB

A typical query is NOT a search, but a complex data transformation
Bringing computation to the lab bench

- Point & Click user interface
- Hide the math, highlight and enable choices
- Large-scale computing resources available on demand
- No software installation/maintenance
- No IT infrastructure
  - No need for sys-admin, obsolescence, space, cooling, etc.

Empowering the stakeholders

- Enable experimental biologists to interactively explore their data
  - Alternative analysis methods
  - Alternative parameter choices
  - Alternative views of results/data
- Relieve computational biologists from non-specialist tasks
  - Allow focus on strengths, e.g. method

Scalability

- Number of users (jobs being run)
- Number/size of scripts and datasets
- Size of individual jobs (CPU + memory required)
- Cost
  - Processor Queues auto-scale with demand
  - Users can have private nodes and shared nodes

Evolvability

- Diversity of users: freely available to all, open source
- Diversity of content: data & scripts provided by the community
  - Community curated
- Inclusive architecture supports
  - multiple OS (Windows, Unix/Linux, Mac)
  - multiple languages (Perl, Python, Java, coming)
Manage script files and directories

Scripts:

create

view

Source code:

```R
# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #
```
Create & delete processing queues

Manage Jobs Queues

Use the interactive interface below to manage your Jobs Queues.

Manage Processing Nodes

Use the interactive interface below to manage your Processing Nodes.
Choose memory and processing node type

**THE INCIDENCE OF ALKAPTONURIA: A STUDY IN CHEMICAL INDIVIDUALITY**

ARCHIBALD E. GARROD

Physician to the Hospital for Sick Children, Great Ormond Street, Demonstrator of Chemical Pathology at St. Bartholomew’s Hospital

Eosinophil counts $\left( \frac{X - \mu}{\sigma} \right)$ per unit volume in 4,458 Icelandic people


Proportion of individuals

% TH1 (INF-γ expressing) cells among CD4+ cells in blood

Duramad et al, Cancer Epidemiology, Biomarkers and Prevention, 2004, 13(9):1452–8


143 healthy middle-aged blood donors


C-reactive protein concentration in blood (mg/L)
2,896 environmental interactions affect 151 out of 165 of Wnt-pathway genes

(395 substances affect Akt1 expression. The graph below excludes Akt1 and the substances that affect it & other genes.)

Of 165 Wnt-pathway associated genes, only 14 have no associations in TCD:

Comparative Toxicogenomics Database [http://ctd.mdibl.org/]
The human microbiome variation


Clinical assessment incorporating a personal genome
(Stephen Quake’s genome)
www.thelancet.com Vol 375 May 1, 2010
DNA sequence variations affecting cellular signaling genes in two individuals
(W=Watson, V=Venter)

<table>
<thead>
<tr>
<th></th>
<th>frameshift_W</th>
<th>frameshift_V</th>
<th>nonsynonymous_W</th>
<th>nonsynonymou_V</th>
<th>UTR_W</th>
<th>UTR_V</th>
<th>splice_site_W</th>
<th>splice_site_V</th>
<th>No. genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCR</td>
<td>0</td>
<td>0</td>
<td>55</td>
<td>52</td>
<td>120</td>
<td>117</td>
<td>2</td>
<td>2</td>
<td>75</td>
</tr>
<tr>
<td>Ca</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>9</td>
<td>182</td>
<td>180</td>
<td>1</td>
<td>1</td>
<td>113</td>
</tr>
<tr>
<td>ErbB</td>
<td>3</td>
<td>3</td>
<td>29</td>
<td>24</td>
<td>161</td>
<td>187</td>
<td>0</td>
<td>1</td>
<td>67</td>
</tr>
<tr>
<td>FR</td>
<td>0</td>
<td>0</td>
<td>29</td>
<td>22</td>
<td>75</td>
<td>82</td>
<td>0</td>
<td>1</td>
<td>59</td>
</tr>
<tr>
<td>JAK2</td>
<td>0</td>
<td>0</td>
<td>77</td>
<td>77</td>
<td>232</td>
<td>232</td>
<td>0</td>
<td>2</td>
<td>156</td>
</tr>
<tr>
<td>MAPK</td>
<td>0</td>
<td>0</td>
<td>77</td>
<td>77</td>
<td>147</td>
<td>147</td>
<td>0</td>
<td>0</td>
<td>294</td>
</tr>
<tr>
<td>mTOR</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>18</td>
<td>94</td>
<td>110</td>
<td>0</td>
<td>4</td>
<td>52</td>
</tr>
<tr>
<td>Notch</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>18</td>
<td>94</td>
<td>96</td>
<td>0</td>
<td>1</td>
<td>43</td>
</tr>
<tr>
<td>PI</td>
<td>3</td>
<td>3</td>
<td>55</td>
<td>55</td>
<td>159</td>
<td>179</td>
<td>0</td>
<td>0</td>
<td>76</td>
</tr>
<tr>
<td>TCR</td>
<td>0</td>
<td>0</td>
<td>27</td>
<td>28</td>
<td>85</td>
<td>139</td>
<td>0</td>
<td>1</td>
<td>87</td>
</tr>
<tr>
<td>TGFβ</td>
<td>0</td>
<td>0</td>
<td>43</td>
<td>39</td>
<td>126</td>
<td>139</td>
<td>12</td>
<td>8</td>
<td>101</td>
</tr>
<tr>
<td>TLR</td>
<td>0</td>
<td>0</td>
<td>43</td>
<td>39</td>
<td>126</td>
<td>139</td>
<td>12</td>
<td>8</td>
<td>101</td>
</tr>
<tr>
<td>VEGF</td>
<td>0</td>
<td>0</td>
<td>50</td>
<td>47</td>
<td>71</td>
<td>117</td>
<td>0</td>
<td>0</td>
<td>76</td>
</tr>
<tr>
<td>Wnt</td>
<td>0</td>
<td>1</td>
<td>55</td>
<td>54</td>
<td>278</td>
<td>246</td>
<td>0</td>
<td>2</td>
<td>106</td>
</tr>
</tbody>
</table>

Abbreviations:
- BCR=B-cell receptor,
- Ca=Calcium,
- ErbB=the v-erb-b2 erythroblastic leukemia viral oncogene family (Receptor Tyrosine Kinases),
- Hh=Hedgehog, PI=Phosphatidylinositol,
- TCR=T-cell receptor,
- TLR=Toll-like receptor.

DNA sequence variations affecting cellular signaling genes in two individuals

![DNA sequence variations affecting cellular signaling genes in two individuals](image-url)
To predict an individual's genomic susceptibilities, we will need to integrate data from:

- Genome annotation DBs
- RNA structure & function DBs
- Protein structure & function DBs
- Pathway structure & function DBs
- Mutation effect prediction algorithms
- Environmental effects DBs
- Drug effects and interactions DBs
- Electronic health records
- Pathway-based calculation of interaction effects
- ...

What is the appropriate computational infrastructure?

**Summary**
Empowering individuals, labs, and consortia
An analysis portal for (all) public data
A place to explore and replicate published work
A place to compare quality of available data & methods

**To come**
Support for multiple languages & Clouds
Strong data encryption for individual genomes
Interactive graphics
Semantic searches
Thanks to:

NHLBI Grant: HL089102

Rajiv Dulepet
Michael Angerman
(Constantin Georgescu)

Ellen Rothenberg & lab
Barbara Wold & lab
(Caltech)

Guy Naor, Cosmin Andriescu
(SparkTech Soft)

Martin Morgan (FHCRC)